NCER National Centre for Econometric Research

NCER Working Paper Series

Are combination forecasts of S&P 500 volatility statistically superior?

Ralf Becker and Adam Clements

Working Paper #17

June 2007

Are combination forecasts of S&P 500 volatility statistically superior?

Ralf Becker* and Adam E. Clements#
* Economics, School of Social Sciences, University of Manchester
# School of Economics and Finance, Queensland University of Technology

May 17, 2007

Abstract

Forecasting volatility has received a great deal of research attention. Many articles have considered the relative performance of econometric model based and option implied volatility forecasts. While many studies have found that implied volatility is the preferred approach, a number of issues remain unresolved. One issue being the relative merit of combination forecasts. By utilising recent econometric advances, this paper considers whether combination forecasts of S&P 500 volatility are statistically superior to a wide range of model based forecasts and implied volatility. It is found that combination forecasts are the dominant approach, indicating that the VIX cannot simply be viewed as a combination of various model based forecasts.

Keywords: Implied volatility, volatility forecasts, volatility models, realized volatility, combination forecasts.
JEL Classification: C12, C22, G00.
Acknowledgements
Corresponding author
Adam Clements
School of Economics and Finance
Queensland University of Technology
GPO Box 2434, Brisbane Q, Australia 4001

email a.clements@qut.edu.au
Ph +61 7 3864 2525, Fax +61 7 3864 1500.

1 Introduction

Estimates of the future volatility of asset returns are of great interest to many financial market participants. Generally, there are two approaches which can be employed to obtain such estimates. *First*, predictions of future volatility can be generated from econometric models of volatility given historical information (model based forecasts, MBF). For surveys of common modeling techniques see Campbell, Lo and MacKinlay (1997) and Gourieroux and Jasiak (2001). *Second*, estimates of future volatility can be derived from option prices using implied volatility (IV). IV should represent a market's best prediction of an assets' future volatility (see, amongst others, Jorion, 1995, Poon and Granger, 2003, 2005).

Given the importance of volatility forecasting, a large number of studies have examined the forecast performance of various approaches. Poon and Granger (2003, 2005) survey the results of 93 articles that consider tests of volatility forecast performance. The general result of this survey was that IV estimates often provide more accurate volatility forecasts than competing MBF. This result is rationalised on the basis that IV should be based on a larger and timelier information set. In a related yet different context Becker, Clements and White (2006) examine whether a particular implied volatility index derived from S&P 500 option prices, the VIX, contains any information relevant to future volatility beyond that reflected in model based forecasts. As they conclude that the VIX does not contain any such information this result, at first sight, appears to contradict the previous findings summarised in Poon and Granger

(2003). However, no forecast comparison is undertaken in Becker, Clements and White (2006) and they merely conjecture that the VIX may be viewed as a combination of MBF.

This paper seeks to examine this contention in more detail, specifically examining the forecast performance of S&P 500 IV, relative to a range of MBF and combination forecasts based on both classes (IV and MBF). In doing so, this paper addresses two outstanding issues raised by Poon and Granger (2003). Poon and Granger (2003) highlight the fact that little attention has been paid to the performance of combination forecasts, which are potentially useful as different forecasting approaches capture different volatility dynamics. They also point out that little has been done to consider whether forecasting approaches are significantly different in terms of performance. By applying the model confidence set approach proposed by Hansen, Lunde and Nason (2003), this paper will determine whether combination volatility forecasts are statistically superior to individual models based and implied volatility forecasts. In doing so, this paper also readdresses the relative performance of IV forecasts.

The paper will proceed as follows. Section 2 will outline the data relevant to this study. Section 3 discusses the econometric models used to generate the various forecasts, along with the methods used to discriminate between forecast performance. Sections 4 and 5 present the empirical results and concluding remarks respectively.

2 Data

This study is based upon data relating to the S&P 500 Composite Index, from 2 January 1990 to 17 October 2003 (3481 observations). To relate to the results of Becker, Clements and White (2006), the same sample period is considered here. To address the research question at hand, estimates of both IV and future actual volatility are required.

The VIX index constructed by the Chicago Board of Options Exchange from S&P 500 index options constitutes the estimate of IV utilised in this paper. It is derived from out-of-the-money put and call options that have maturities close to the target of 22 trading days. For technical details relating to the construction of the VIX index, see Chicago Board Options Exchange (CBOE, 2003). While the true process underlying option pricing is unknown, the VIX is constructed to be a general measure of the market's estimate of average S&P 500 volatility over the subsequent 22 trading days (BPT, 2001, Christensen and Prabhala, 1998 and CBOE, 2003). Having a fixed forecast horizon is advantageous and avoids various econometric issues. This index has only been available since September 2003 when the CBOE replaced a previous implied volatility index based on S&P 100 options[1]. It's advantages in comparison to the previous implied volatility index is that it no longer relies on option implied volatilities derived from Black-Scholes option pricing models, it is based on more liquid options written on the S&P500 and is easier to hedge against (CBOE, 2003).

[1]The new version of the VIX has been calculated retrospectively back to January 1990, the beginning of the sample considered here.

For the purposes of this study estimates of actual volatility were obtained using the realized volatility (RV) methodology outlined in ABDL (2001, 2003). RV estimates volatility by means of aggregating intra-day squared returns. It should be noted that the daily trading period of the S&P500 is 6.5 hours and that overnight returns were used as the first intra-day return in order to capture the variation over the full calender day. ABDL (1999) suggest how to deal with practical issues relating to intra-day seasonality and sampling frequency when dealing with intra-day data. Based on this methodology, daily RV estimates were constructed using 30 minute S&P500 index returns[2]. It is widely acknowledged (see e.g. Poon and Granger, 2003) that RV is a more accurate and less noisy estimate of the unobservable volatility process than squared daily returns. Patton (2006) suggests that this property of RV is beneficial when RV is used a proxy for observed volatility when evaluating forecasts.

Figure 1 shows the VIX and daily S&P500 RV for the sample period considered. While the RV estimates exhibit a similar overall pattern when compared to the VIX, RV reaches higher peaks than the VIX. This difference is mainly due to the fact that the VIX represents an average volatility measure for a 22 trading day period as opposed to RV that is a measure of daily volatility.

3 Methodology

In this section the econometric models upon which forecasts are based will be outlined, followed by how the competing forecasts will be combined. This section concludes with a discussion of the technique utilised to discriminate

[2]Intraday S&P 500 index data were purchased from Tick Data, Inc.

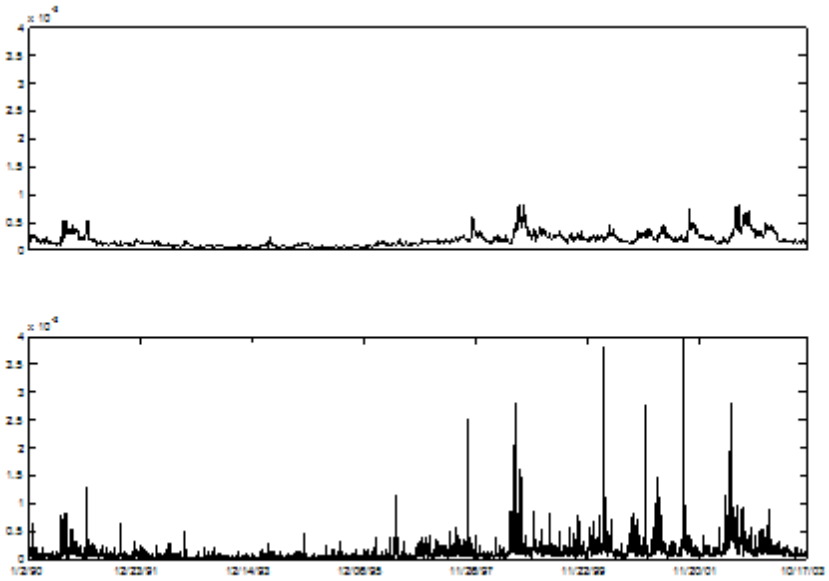

Figure 1: Daily VIX index (top panel) and daily S&P 500 index RV estimate (bottom panel).

between the volatility forecasts.

3.1 Model based forecasts

While the true process underlying the evolution of volatility is not known, a range of candidate models exist and are chosen so that they span the space of available model classes. The set of models chosen are based on the models considered when the informational content of IV has been considered in Koopman, Jungbacker and Hol (2004) and BPT (2001) and Becker, Clements and White (2006). The models chosen include models from the GARCH, Stochastic volatility (SV), and RV classes. A brief summary of the in-sample fit of the models will be given in this section[3].

GARCH style models employed in this study are similar to those proposed

[3]For more detailed estimation results see Becker, Clements and White (2006).

by BPT (2001). The simplest model specification is the GJR (see Glosten, Jagannathan and Runkle, 1993, Engle and Ng, 1991) process,

$$\begin{aligned} r_t &= \mu + \varepsilon_t \qquad \varepsilon_t = \sqrt{h_t} z_t \qquad z_t \sim N(0,1) \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 s_{t-1} \varepsilon_{t-1}^2 + \beta h_{t-1} \end{aligned} \tag{1}$$

that captures the asymmetric relationship between volatility and returns, with s_{t-1} taking the value of unity when $\varepsilon_{t-1} < 0$ and 0 otherwise. This process nests the standard GARCH model when $\alpha_2 = 0$.

Following BPT (2001), standard GARCH style models are augmented by the inclusion of RV[4]. The most general specification of a GARCH process including RV is given by,

$$\begin{aligned} r_t &= \mu + \varepsilon_t \qquad \varepsilon_t = \sqrt{h_t} z_t \qquad z_t \sim N(0,1) \\ h_t &= h_{1t} + h_{2t} \\ h_{1t} &= \alpha_0 + \beta h_{t-1} + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 s_{t-1} \varepsilon_{t-1}^2 \\ h_{2t} &= \gamma_1 h_{2t-1} + \gamma_2 RV_{t-1} \end{aligned} \tag{2}$$

and is defined as the GJRRVG model. This allows for two components to contribute to volatility, with each component potentially exhibiting persistence. This specification nest various other models, GJRRV if $\gamma_1 = 0$, GARCHRV if $\gamma_1 = \alpha_2 = 0$, GJR if $\gamma_1 = \gamma_2 = 0$ and GARCH if $\gamma_1 = \gamma_2 = \alpha_2 = 0$.

Parameter estimates for the GARCH and GJR models are similar to those commonly observed for GARCH models based on various financial time series,

[4]While BPT (2001) also extend the GJR model to include the VIX index, this is not relevant to the current study as it is the goal to seperate forecast performance of IV and MBF.

reflecting strong volatility persistence, and are qualitatively similar to those reported in BPT (2001)[5]. Furthermore, allowing for asymmetries in conditional volatility is important, irrespective of the volatility process considered.

While not considered by BPT 2001, this study also proposes that an SV model may be used to generate forecasts. SV models differ from GARCH models in that conditional volatility is treated as an unobserved variable, and not as a deterministic function of lagged returns. The simplest SV model describes returns as

$$r_t = \mu + \sigma_t \, u_t \qquad u_t \sim N\left(0,1\right) \tag{3}$$

where σ_t is the time t conditional standard deviation of r_t. SV models treats σ_t as an unobserved (latent) variable, following its own stochastic path, the simplest being an AR(1) process,

$$\log\left(\sigma_t^2\right) = \alpha + \beta \log\left(\sigma_{t-1}^2\right) + w_t \qquad w_t \sim N(0, \sigma_w^2). \tag{4}$$

Similar to Koopman *et al.* (2004), this study extends a standard volatility model to incorporate RV as an exogenous variable in the volatility equation. The standard SV process in equation (4) can be extended to incorporate RV in the following manner and is denoted by SVRV

$$\log\left(\sigma_t^2\right) = \alpha + \beta \, \log\left(\sigma_{t-1}^2\right) + \gamma(\log(RV_{t-1}) - E_{t-1}[\log\left(\sigma_{t-1}^2\right)]) + w_t. \tag{5}$$

Here, RV enters the volatility equation through the term $\log(RV_{t-1}) - E_{t-1}$ $[\log\left(\sigma_{t-1}^2\right)]$. This form is chosen due to the high degree of correlation between

[5] As the models discussed in this section will be used to generate 2,460 recursive volatility forecasts (see Section 3) reporting parameter estimates is of little value. Here we will merely discuss the estimated model properties qualitatively. Parameter estimates for the recursive windows and the full sample are available on request.

RV and the latent volatility process and represents the incremental information contained in the RV series. It is noted that equation (5) nests the standard SV model as a special case by imposing the restriction $\gamma = 0$.[6] The SV models appear to capture the same properties of the volatility process as the GARCH-type models. In both instances, volatility is found to be a persistent process, and the inclusion of RV as an exogenous variable is important.

In addition to GARCH and SV approaches, it is possible to utilise estimates of RV to generate forecasts of future volatility. These forecasts can be generated by directly applying time series models, both short and long memory, to daily measures of RV, RV_t. In following ADBL (2003) and Koopman *et al.* (2004) ARMA(2,1) and ARFIMA(1,d,0) processes are utilised. In its most general form an ARFIMA(p,d,q) process may be represented as

$$A(L)\,(1-L)^d\left(x_t - \mu_{x_t}\right) = B(L)\,\varepsilon_t, \tag{6}$$

where $A(L)$ and $B(L)$ are coefficient polynomials of order p and q. The degree of fractional integration is d. A general ARMA(p,q) process applied to x_t is defined under the restriction of $d = 0$. In the context of this work ARMA(2,1) and ARFIMA(1,d,0) were estimated with $x_t = \sqrt{RV_t}$ and $x_t = \ln\left(\sqrt{RV_t}\right)$ [**RB: Which one doe we use??**]. These variable transformations are applied to reduce the skewness and kurtosis of the observed volatility data (Andersen *et al.*, 2003). In the ARMA (2,1) case, parameter estimates reflect the common feature of volatility persistence. Allowing for fractional integration in the

[6] Numerous estimation techniques may be applied to the model in equations 3 and 4 or 5. In this instance the nonlinear filtering approach proposed by Clements, Hurn and White (2003) is employed. This approach is adopted as it easily accommodates exogenous variables in the state equation.

ARFIMA(1,d,0) case reveals that volatility exhibits long memory properties.

In order to generate model based volatility forecasts which capture the information available at time t efficiently, the volatility models were reestimated for time-step t using data from $t-999$ to t. The resulting parameter values were then used to generate 22 day-ahead volatility forecasts ($t+1 \rightarrow t+22$), this time horizon is used for comparability with the VIX IV forecast. The first forecast period covers the trading period from 13 December 1993 to 12 January 1994. For subsequent forecasts the model parameters were reestimated using a sliding estimation window of 1000 observations. The last forecast period covers 18 September 2003 to 17 October 2003, leaving 2460 forecasts.

3.2 Combining forecasts

Two strategies have been employed to construct combination forecasts. The simplest, and most naïve approach sets the combinations to be the mean of the constituent forecasts, thus an equally weighted combination of each forecast.

The alternative is to utilise the regression combination approach discussed in Clements and Hendry (1998) where the combination weights are derived from the following regression,

$$\overline{RV}_{t+22} = \alpha_0 + \alpha_1 \ f_t^1 + \alpha_2 \ f_t^2 + ... + \alpha_n \ f_t^n + \epsilon_t \tag{7}$$

where $\overline{RV}_{t+22}$ is the target volatility, the average RV over the 22 day forecast horizon ($t+1$ to $t+22$) and $f_t^i, \ i = 1, \ 2, ..., n$ are n different forecasts of average volatility ($t+1$ to $t+22$) formed at time t to be included in the combination. The resulting combination forecast is then given by

$$f_t^c = \widehat{\alpha}_0 + \widehat{\alpha}_1 \ f_t^1 + \widehat{\alpha}_2 \ f_t^2 + ... + \widehat{\alpha}_n \ f_t^n \tag{8}$$

where f_t^c is the combination forecast. In this context, the combination parameters have been estimated based on a rolling window of forecasts. A fixed set of weights was deemed to be inappropriate in this context as the level of volatility was substantially higher during the latter part of the sample period, see Figure 1. Therefore, to form a combination forecast at time t, f_t^c, combination weights were obtained by estimating equation 7 on forecasts from $t-500$ to $t-1$, and then combining the various individual forecasts f_t^i formed at time t using these weights. Allowing for the initial period of 500 forecasts to be used for estimation, the final 1960 forecasts are used for comparative purposes. The specific composition of the combination forecasts will be discussed in Section 4 as they are motivated by results based on the individual forecasts.

3.3 Evaluating forecasts

As argued above, it is the objective of this paper to determine whether combination forecasts are superior to individual MBF and IV. At the heart of the model confidence set (MCS) methodology (Hansen, Lunde and Nason, 2003) as it is applied here, is a forecast loss measure. Such measures have frequently been used to rank different forecasts and the two loss functions utilised here are the, MSE and QLIKE,

$$MSE^i = (\overline{RV}_{t+22} - f_t^i)^2, \tag{9}$$

$$QLIKE^i = \log(f_t^i) + \frac{\overline{RV}_{t+22}}{f_t^i}, \tag{10}$$

where f_t^i are individual forecasts obtained from individual models (and the VIX) along with combination forecasts based on equal weights or regression weights. The total number of candidate forecasts will be denoted as m_0, therefore the competing forecasts, individual and combination are given by $f_t^i,\ i = 1,\ 2, ..., m_0$. While there are many alternative loss functions, Patton (2006) shows that MSE and QLIKE belong to a family of loss functions that are robust to noise in the volatility proxy, $\overline{RV}_{t+22}$ in this case. Each loss function has somewhat different properties, MSE is symmetric whereas QLIKE penalises under-prediction more heavily than over-prediction.

While these loss functions allow forecasts to be ranked, they give no indication of whether the performance of the forecasts are significantly different. The model confidence set (MCS) approach allows for such conclusions to be drawn. The interpretation attached to an MCS is that it contains the best forecast with a given level of confidence. The MCS may contain a number of models which indicates they are of equal predictive ability (EPA). The construction of an MCS is an iterative procedure in that it requires a sequence of tests for EPA. The set of candidate models is trimmed by deleting models that are found to be inferior. The final surviving set of models in the MCS contain the optimal model with a given level of confidence and are not significantly different in terms of their forecast performance.

The procedure starts with a full set of candidate models $\mathcal{M}_0 = \{1, ..., m_0\}$. The MCS is determined by sequentially trimming models from $\mathcal{M}_0$ therefore reducing the number of models to $m < m_0$. Prior to starting the sequential elimination procedure, all loss differentials between models i and j are com-

puted,

$$d_{ij,t} = L(\overline{RV}_{t+22}, f_t^i) - L(\overline{RV}_{t+22}, f_t^j), \qquad i,j = 1,...,m_0, \qquad t = 1,...,T \tag{11}$$

where $L()$ is chosen to be one of the loss functions described above. At each step, the EPA hypothesis

$$H_0 \ : \ E(d_{ij,t}) = 0, \qquad \forall \ i > j \in \mathcal{M} \tag{12}$$

is tested for a set of models $\mathcal{M} \subset \mathcal{M}_0$, with $\mathcal{M} = \mathcal{M}_0$ at the initial step. If H_0 is rejected at the significance level α, the worst performing model is removed and the process continued until non-rejection occurs with the set of surviving models being the MCS, $\widehat{\mathcal{M}}^*_\alpha$. If a fixed significance level α is used at each step, $\widehat{\mathcal{M}}^*_\alpha$ contains the best model from $\mathcal{M}_0$ with $(1-\alpha)$ confidence[7].

At the core of the EPA statistic is the t-statistic

$$t_{ij} = \frac{\overline{d}_{ij}}{\sqrt{\widehat{var}(\overline{d}_{ij})}}$$

where $\overline{d}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}$. t_{ij} provides scaled information on the average difference in the forecast quality of models i and j. $\widehat{var}(\overline{d}_{ij})$ is an estimate of $var(\overline{d}_{ij})$ and is obtained from a bootstrap procedure described below. In order to decide whether, at any stage, the MCS must be further reduced, the null hypothesis in (12) is to be evaluated. The difficulty being that for each set $\mathcal{M}$ the information from $(m-1)\,m/2$ unique t-statistics needs to be distilled into one test statistic. Hansen, et al. (2003, 2005) propose the following the range statistic,

[7] See Hansen *et al.* (2005) for a discussion of this interpretation.

$$T_R = \max_{i,j \in \mathcal{M}} |t_{ij}| = \max_{i,j \in \mathcal{M}} \frac{|\overline{d}_{ij}|}{\sqrt{\widehat{var}(\overline{d}_{ij})}} \tag{13}$$

and a semi-quadratic statistic,

$$T_{SQ} = \sum_{\substack{i,j \in \mathcal{M} \\ i<j}} t_{ij}^2 = \sum_{\substack{i,j \in \mathcal{M} \\ i<j}} \frac{(\overline{d}_{ij})^2}{\widehat{var}(\overline{d}_{ij})} \tag{14}$$

as test statistics to establish EPA. Both test statistics indicate a rejection of the EPA hypothesis for large values. The actual distribution of the test statistic is complicated and depends on the structure between the forecasts included in $\mathcal{M}$. Therefore p-values for each of these test statistics have to be obtained from a bootstrap distribution (see below). When the null hypothesis of EPA is rejected, the worst performing model is removed from $\mathcal{M}$. The latter is identified as $\mathcal{M}_i$ where

$$i = \arg \max_{i \in \mathcal{M}} \frac{\overline{d}_i}{\sqrt{\widehat{var}(\overline{d}_{i.})}} \tag{15}$$

and $\overline{d}_{i.} = \frac{1}{m-1} \sum_{j \in \mathcal{M}} \overline{d}_{ij}$. The tests for EPA are then conducted on the reduced set of models and one continues to iterate until the null hypothesis of EPA is not rejected.

Bootstrap distributions are required for the test statistics T_R and T_{SQ}. These distributions will be used to estimate p-values for T_R and T_{SQ} tests and hence calculate model specific p-values. At the core of the bootstrap procedure is the generation of bootstrap replications of $d_{ij,t}$. In doing so, the temporal dependence in $d_{ij,t}$ must be accounted for. This is achieved by the block bootstrap, which is conditioned on the assumption that the $\{d_{ij,t}\}$ sequence is stationary and follows a strong geometric mixing assumption. The basic steps of the bootstrap procedure are now described.

Let $\{d_{ij,t}\}$ be the sequence of T observed differences in loss functions for models i and j. B block bootstrap counterparts are generated for all combinations of i and j, $\left\{d_{ij,t}^{(b)}\right\}$ for $b = 1, ..., B$. Values with a bar, e.g. $\overline{d}_{ij} = T^{-1}\sum d_{ij,t}$, represent averages over all T observations. First we will establish how to estimate the variance estimates $\widehat{var}(\overline{d}_{ij})$ and $\widehat{var}(\overline{d}_{i.})$, which are required for the calculation of the EPA test statistics in (13), (14) and (15):

$$\begin{aligned}\widehat{var}(\overline{d}_{ij}) &= B^{-1}\sum_{b=1}^{B}\left(\overline{d}_{ij}^{(b)} - \overline{d}_{ij}\right)^2 \\ \widehat{var}(\overline{d}_{i.}) &= B^{-1}\sum_{b=1}^{B}\left(\overline{d}_{i.}^{(b)} - \overline{d}_{i.}\right)^2\end{aligned}$$

for all $i, j \in \mathcal{M}$. In order to evaluate the significance of the EPA test a p-value is required. That is obtained by comparing T_R or T_{SQ} with bootstrap realisations $T_R^{(b)}$ or $T_{SQ}^{(b)}$.

$$\begin{aligned}\widehat{p}_\tau &= B^{-1}\sum_{b=1}^{B} 1\left(T_\tau^{(b)} > T_\tau\right) \quad \text{for } \tau = R, SQ \\ 1(A) &= \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}.\end{aligned}$$

The B bootstrap versions of the test statistics T_R or T_{SQ} are calculated by replacing $\left|\overline{d}_{ij}\right|$and $\left(\overline{d}_{ij}\right)^2$ in equations (13) and (14) with $\left|\overline{d}_{ij}^{(b)} - \overline{d}_{ij}\right|$ and $\left(\overline{d}_{ij}^{(b)} - \overline{d}_{ij}\right)^2$ respectively. The denominator in the test statistics remains the bootstrap estimate discussed above.

This model elimination process can be used to produce model specific p-values. A model is only accepted into $\widehat{\mathcal{M}}_\alpha^*$ if its p-value exceeds α. Due to the definition of $\widehat{\mathcal{M}}_\alpha^*$ this implies that a model which is not accepted into $\widehat{\mathcal{M}}_\alpha^*$ is unlikely to belong to the set of best forecast models. The model specific p-values are obtained from the p-values for the EPA tests described above. As the kth

model is eliminated from $\mathcal{M}$, save the (bootstrapped) p-value of the EPA test in (13) or (14) as $p(k)$. For instance, if model $\mathcal{M}_i$ was eliminated in the third iteration, i.e. $k = 3$. The p-value for this ith model is then $\widehat{p}_i = \max_{k \leq 3} p(k)$. This ensures that the model eliminated first is associated with the smallest p-value indicating that it is the least likely to belong into the MCS[8].

4 Empirical results

Results pertaining to individual forecasts, including the VIX and MBF will be discussed first, followed by those including the combination forecasts. The exact composition of the combinations will be outlined once the individual forecasts are compared as their composition is motivated by the performance of the individual forecasts. Rankings based simply on the loss functions, MSE and $QLIKE$ will be discussed first followed by an examination of the MCS.

4.1 Individual forecasts

Table 1 reports the ranking based on both MSE and QLIKE for all of the individual forecasts. The rankings given the two different loss functions differ slightly, as they penalise forecast errors differently. A number of interesting patterns emerge. The ARMA and ARFIMA time series forecasts based on RV produce the most accurate forecasts of $\overline{RV}_{t+22}$, confirming similar results (e.g. Anderson *et al.*, 2003). Another obvious result is that VIX[9] is not amongst the most accurate forecasts under either loss function, although it does better under the asymmetric loss function. Further it is apparent that the GARCH models

[8] See Hansen et al. (2005) for a detailed interpretation for the MCS p-values.

[9] Poon and Granger (2003) suggest to divide the VIX by $\sqrt{365/252}$ to account for the difference between calender month and trading days.

MSE		$QLIKE$	
$ARMA$	1.659	$ARFIMA$	3159.2
$ARFIMA$	1.667	$ARMA$	3174.3
$GARCHRV$	1.952	$SVRV$	3222.5
$GJRRV$	2.161	VIX	3253.0
$GJRRVG$	2.404	$GARCHRV$	3266.6
VIX	2.525	$GJRRV$	3278.7
$GARCH$	2.575	$GARCH$	3472.5
SV	2.730	GJR	3575.7
GJR	2.857	$GJRRVG$	3700.7
$SVRV$	4.543	SV	3923.0

Table 1: Loss function rankings for individual forecasts.

$Model$	T_R $\widehat{p}$	MCS $\widehat{p}_i$	T_{SQ} $\widehat{p}$	MCS $\widehat{p}_i$
$SVRV$	0.013	0.013	0.005	0.005
GJR	0.016	0.016	0.005	0.005
SV	0.016	0.016	0.002	0.005
$GARCH$	0.011	0.016	0.003	0.005
VIX	0.011	0.016	0.004	0.005
$GJRRVG$	0.009	0.016	0.000	0.005
$GJRRV$	0.005	0.016	0.000	0.005
$GARCHRV$	0.008	0.016	0.006	0.006
$ARFIMA$	0.844	0.844	0.844	0.844
$ARMA$	–	1.000	–	1.000

Table 2: MCS results for individual forecasts given the MSE loss function. The first row respresents the first model removed, down to the best performing model in the last row.

that incorporate RV measures into the volatility equation are more accurate out-of-sample than those that do not.

MCS results will now reveal whether the VIX is significantly inferior to those forecasts with lower average loss.

Table 2 reports the MCS results for the individual forecasts based on the MSE loss function. It turns out that the model with the largest T_R test statistic is the $SVRV$ model. The p-value, determined in the first reduction round is 0.013. As it is eliminated in the first round this automatically determines the

MCS p-value for $SVRV$ to be 0.013. In the second round of elimination the GJR model fares worst. It produces the largest T_R test statistic (p-value of 0.016) and is therefore dropped at a 5% significance level. As the p-value of 0.016 is larger than the MCS p-values of the model(s) previously dropped this is also its MCS p-value. In fact, at a 5% significance level 6 more models are dropped from $\mathcal{M}$ and only the $ARFIMA$ and $ARMA$ survive in the MCS and therefore constitute $\widehat{\mathcal{M}}^*_{0.05}$. As can be seen from the last two columns this result does not change qualitatively if one considers the T_{SQ} statistic rather than the T_R statistic.

Therefore ARFIMA and ARMA constitute the MCS with 95% confidence and are significantly superior to other competing models and the VIX. Results based on the QLIKE loss function, reported in Table 3, reveals somewhat of a different picture. At a 5% significance, the MCS contains $SVRV$, VIX, $ARFIMA$ and $ARMA$. Under the QLIKE loss assumption, the VIX is of EPA but clearly not superior to the three surviving MBF. As the asymmetric loss function admits two additional models in the MCS it can be conjectured that the $SVRV$ and VIX in fact avoid significantly underpredicting volatility, as that is the mistake most heavily penalised under this loss function.

These results will motivate the combinations forecasts formed in the following section. The results of the following section will reveal whether the VIX is of EPA relative to these combinations and can therefore be viewed as a combination of various MBF.

$Model$	T_R $\widehat{p}$	MCS $\widehat{p}_i$	T_{SQ} $\widehat{p}$	MCS $\widehat{p}_i$
SV	0.000	0.000	0.000	0.000
$GJRRVG$	0.000	0.000	0.000	0.000
GJR	0.000	0.000	0.000	0.000
$GARCH$	0.000	0.000	0.001	0.001
$GJRRV$	0.002	0.002	0.004	0.004
$GARCHRV$	0.007	0.007	0.032	0.032
VIX	0.200	0.200	0.150	0.150
$SVRV$	0.134	0.200	0.100	0.150
$ARMA$	0.219	0.219	0.219	0.219
$ARFIMA$	–	1.000	–	1.000

Table 3: MCS results for individual forecasts given the QLIKE loss function. The first row respresents the first model removed, down to the best performing model in the last row.

4.2 Combination forecasts

Based on the MCS results for the individual forecasts, the first set of combination forecasts are chosen. Therefore combinations of ARMA+ARFIMA and ARMA+ARFIMA+SVRV+VIX were formed given the MSE and QLIKE results reported above. The final two sets of combinations are natural to consider, a combination of all MBF and all forecasts, denoted as ALLMBF and ALL in this section respectively. As discussed in Section 3, the combination forecasts are constructed using both a simple average or regression weighted function of the constituent forecasts. Both of these approaches are applied to each of the combinations described here with the simple average and regression based combinations indicated by u and r superscripts respectively. In total, the performance of 19 forecasts are compared in this section, 11 individual and 8 combination forecasts.

Table 4 ranks all 19 forecasts based on both loss functions. It is evident that the most accurate forecasts, irrespective of the loss function are the combina-

MSE		$QLIKE$	
$ARMA+ARFIMA^r$	1.477	$\binom{ARMA+ARFIMA}{+SVRV+VIX}^r$	3042.5
$\binom{ARMA+ARFIMA}{+SVRV+VIX}^r$	1.502	$ARMA+ARFIMA^r$	3043.8
ALL^r	1.509	$\binom{ARMA+ARFIMA}{+SVRV+VIX}^u$	3102.4
$ALLMBF^r$	1.549	$ARFIMA$	3159.2
$ARMA+ARFIMA^u$	1.648	$ARMA+ARFIMA^u$	3161.6
$ARMA$	1.659	$ARMA$	3174.3
$ARFIMA$	1.667	ALL^u	3205.2
$GARCHRV$	1.952	$SVRV$	3222.5
$\binom{ARMA+ARFIMA}{+SVRV+VIX}^u$	1.969	VIX	3253.0
ALL^u	2.001	$ALLMBF^u$	3266.1
$ALLMBF^u$	2.014	$GARCHRV$	3266.6
$GJRRV$	2.161	$GJRRV$	3278.7
$GJRRVG$	2.404	$ALLMBF^r$	3447.5
VIX	2.525	$GARCH$	3472.5
$GARCH$	2.575	GJR	3575.7
SV	2.730	$GJRRVG$	3700.7
GJR	2.857	SV	3923.0
$SVRV$	4.543	ALL^r	5796.9

Table 4: Loss function rankings for individual and combination forecasts.

tion forecasts based on the MCS of individual forecasts, the regression based combinations of $ARMA$+$ARFIMA$ and $ARMA$+$ARFIMA$+$SVRV$+VIX. The equally weighted combinations of these also rank relatively highly, along with individual $ARMA$ and $ARFIMA$ forecasts. As an individual forecast, the VIX does not perform particularly well. This represents a preliminary indication that the VIX, as an IV estimate, does not seem to incorporate information relevant to future volatility of the same quality as that contained in the top performing combinations. The issue of whether the VIX is significantly inferior will now be considered.

Tables 5 and 6 contain the MCS results given the MSE and QLIKE loss functions respectively. Assuming a level of significance of 5%, the MSE MCS contains predominantly $ARMA$ and $ARFIMA$ based forecasts. While ALLMBF^r

and ALL^r are contained in the MCS, it may be conjectured that the role played by the RV time series ARMA and ARFIMA forecasts is responsible for this as no other individual forecast is included. Once again, the VIX as an individual forecast is significantly inferior to these individual and combination forecasts. The only role played by the VIX in this case is as a constituent of the $ARMA$+$ARFIMA$+$SVRV$+VIX^r and ALL combinations. A similar pattern emerges when the $QLIKE$ results in Table 6 are examined. In this case the MCS is narrower and does not contain the ALL^r or ALLMBF^r forecasts. Once again the VIX forecast is clearly inferior to these combination forecasts. Interestingly, the MCS based on the QLIKE loss function does not include the individual $ARMA$ and $ARFIMA$ models any longer. This indicates that here, in contrast to the case of the MSE loss function, the combination of forecasts delivers a statistically significant advantage.

A number of interesting patterns emerge from these results. From a practical viewpoint, it is clear that combination forecasts have the potential to produce forecasts of superior accuracy relative to the individual forecasts. This is not surprising as different models capture different dynamics in volatility. If the top performing individual forecasts are combined this may lead to a dominant combination forecast, superior to its individual constituents and other competing models. In the present context, however, this is only true for the asymmetric loss function QLIKE.

The results also shed further light on the manner in which IV estimates (VIX in this case) are formed. These results suggest that option traders, when forming volatility forecasts, are not taking into account the same quality

of information that is reflected in the combination (and some of the individual) model based volatility forecasts. These findings extend those of Becker, Clements and White (2006). There it was established that the VIX did not contain any information superior to that in the combination of all MBF. This research goes further and it can now be claimed that the VIX, not only contains no additional information, but also does not fully incorporate the information contained in MBF. The VIX can, therefore, not be seen as the best possible combination of all MBF.

Two interpretations can be attached to this finding. First, it could be argued that the options market is not informationally efficient in the sense that it does not incorporate all information available from MBF of volatility. While this paper finds statistical evidence to support this statement, a more robust check would be to establish whether the use of the statistically superior volatility forecast would deliver significant excess profits in an appropriate trading strategy. Second, it is possible that a time-varying volatility risk premium breaks the link between IV and actual realised volatility[10]. Making allowance for this possibility is beyond the scope of this paper.

5 Conclusion

Issues relating to forecasting volatility have attracted a great deal of attention in recent years. There have been many studies into the relative merits of implied and model based volatility forecasts. It has often been found that implied volatility offers a superior volatility forecast when compared against individual

[10] There is some evidence for time variation in the volatility risk premium (e.g. Bollerslev et al., 2006).

$Model$	T_R $\widehat{p}$	MCS $\widehat{p}_i$	T_{SQ} $\widehat{p}$	MCS $\widehat{p}_i$
GJR	0.033	0.033	0.003	0.003
SV	0.033	0.033	0.011	0.011
VIX	0.022	0.033	0.010	0.011
$SVRV$	0.023	0.033	0.004	0.011
$GARCH$	0.023	0.033	0.010	0.011
$GJRRV$	0.018	0.033	0.018	0.018
$GJRRVG$	0.016	0.033	0.019	0.019
$ALLMBF^u$	0.017	0.033	0.020	0.020
ALL^μ	0.025	0.033	0.035	0.035
$\binom{ARMA+ARFIMA}{+SVRV+VIX}^u$	0.022	0.033	0.066	0.066
$GARCHRV$	0.023	0.033	0.113	0.113
$ARMA$	0.548	0.548	0.513	0.513
$ARFIMA$	0.519	0.548	0.487	0.513
$ARMA + ARFIMA^u$	0.541	0.548	0.624	0.624
$ALLMBF^r$	0.882	0.882	0.803	0.803
$\binom{ARMA+ARFIMA}{+SVRV+VIX}^r$	0.818	0.882	0.851	0.851
ALL^r	0.767	0.882	0.767	0.851
$ARMA + ARFIMA^r$	–	1.000	–	1.000

Table 5: MCS results for individual forecasts given the MSE loss function. The first row respresents the first model removed, down to the best performaing model in the last row.

model based volatility forecasts. This paper has readdressed this question in the context of S&P 500 implied volatility, the VIX index. The forecast performance of the VIX index has been compared to a range of model based forecasts and combination forecasts. In doing so, further light is shed on the nature of the information reflected in the VIX forecast.

In practical terms the VIX index produces forecasts that are inferior to a number of competing model based forecasts, namely time series models of realised volatility. The significance of these differences has been evaluated using the model confidence set technology by Hansen *et al.* (2003, 2005). As it turns out the VIX is not significantly inferior when an asymmetric loss function is used. When the best model based volatility forecasts are combined they are

$Model$	T_R $\widehat{p}$	MCS $\widehat{p}_i$	T_{SQ} $\widehat{p}$	MCS $\widehat{p}_i$
$GJRRVG$	0.000	0.000	0.001	0.001
SV	0.001	0.001	0.001	0.001
GJR	0.000	0.001	0.000	0.001
$GARCH$	0.002	0.002	0.001	0.001
$GJRRV$	0.000	0.002	0.000	0.001
$ALLMBF^u$	0.002	0.002	0.008	0.008
$GARCHRV$	0.007	0.007	0.009	0.009
$SVRV$	0.006	0.007	0.006	0.009
VIX	0.022	0.022	0.034	0.034
ALL^u	0.017	0.022	0.026	0.034
$ARMA$	0.055	0.055	0.047	0.047
ALL^r	0.082	0.082	0.091	0.091
$ARMA + ARFIMA^u$	0.078	0.082	0.066	0.091
$ARFIMA$	0.082	0.082	0.106	0.106
$\binom{ARMA+ARFIMA}{+SVRV+VIX}^u$	0.150	0.150	0.212	0.212
$ALLMBF^r$	0.827	0.827	0.795	0.795
$ARMA + ARFIMA^r$	0.867	0.867	0.867	0.867
$\binom{ARMA+ARFIMA}{+SVRV+VIX}^r$	–	1.000	–	1.000

Table 6: MCS results for individual forecasts given the QLIKE loss function. The first row respresents the first model removed, down to the best performaing model in the last row.

found to be superior to the individual mode based and VIX forecasts. In summary, the most accurate S&P 500 volatility forecast is obtained from a combination of short and long memory models of realised volatility. While previous work has found that the VIX contains no information beyond that contained in model based forecasts. These findings indicate that, while it is entirely plausible that the implied volatility combines information used in a range of different model based forecasts, it is not the best possible combination of such information. When compared to other combined forecasts, the VIX drops out of the model confidence set.

References

Andersen, T.G., and Bollerslev, T., and Diebold, F.X., and Labys, P. (1999) "(Understanding, optimizing, using and forecasting) Realized Volatility and Correlation." Working Paper, University of Pennsylvania.

Andersen T.G., Bollerslev T., Diebold F.X. and Labys P. (2001). "The distribution of exchange rate volatility." Journal of the American Statistical Association 96, 42-55.

Andersen T.G., Bollerslev T., Diebold F.X. and Labys P. (2003). "Modeling and forecasting realized volatility." Econometrica 71, 579-625.

Blair B.J., Poon S-H. and Taylor S.J. (2001). "Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns." Journal of Econometrics 105, 5-26.

Becker, R.and Clements, A. and White, S. (2006) "Does implied volatility provide any information beyond that captured in model-based volatility forecasts?", forthcoming in *Journal of Banking and Finance.*

Bollesrlev T., M. Gibson and H. Zhou (2006). Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities. unpublished manuscript, Duke University.

Campbell, J.Y. and Lo, A.W. and MacKinlay, A.G. (1997). The Econometrics of Financial Markets. Princeton University Press, Princeton NJ.

Chernov M. (2001) "Implied Volatilities as Forecasts of Future Volatility, Time-Varying Risk Premia, and Returns Variability", unpublished manuscript, Columbia University.

Chernov M. (2002) "On the Role of Volatility Risk Premia in Implied Volatilities Based Forecasting Regressions", unpublished manuscript, Columbia University.

Chicago Board of Options Exchange (2003) VIX, CBOE Volatility Index.

Clements M.P. and Hendry D.F., 1998. Forecasting Economic Time Series. Cambridge University Press: Cambridge.

Clements, A.E., Hurn, A.S, White, S.I., 2003. Discretised Non-Linear Filtering of Dynamic Latent Variable Models with Application to Stochastic Volatility. Discussion Paper No 179, School of Economics and Finance, Queensland University of Technology.

Christensen B.J. and Prabhala N.R. (1998). "The relation between implied and realized volatility." Journal of Financial Economics 50, 125-150.

Engle, R.F,.Ng, V.K., 1991. Measuring and testing the impact of news on volatility. Journal of Finance, 48, 1749-1778.

Glosten, L.R., Jagannathan, R,.Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance, 48, 1779-1801.

Gourieroux C. and Jasiak J. (2001). Financial Econometrics. Princeton University Press: Princeton.

Hansen, P.R., Lunde, A. and Nason, J.M. (2003) "Choosing the best volatility models: the model confidence set appraoch", *Oxford Bulletin of Economics and Statistics,* 65, 839-861.

Hansen, P.R., Lunde, A. and Nason, J.M. (2005) "Model confidence sets for forecasting models appraoch", *Working Paper 2005-7,* Brown University.

Jorion P. (1995). "Predicting volatility in the foreign exchange market.", *Journal of Finance,* 50, 507-528.

Koopman, S.J., Jungbacker, B,.Hol, E., 2005. Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. Journal of Empirical Finance, 12, 445-475.

Patton, A.J. (2005) "Volatility forecast comparison using imperfect volatility proxies", Unpublised

Poon S-H. and Granger C.W.J. (2003). "Forecasting volatility in financial markets: a review." *Journal of Economic Literature,* 41, 478-539.

Poon S-H. and Granger C.W.J. (2005). "Practical Issues in forecasting volatility" *Financial Analysts Journal*, 61, 45-56.